DOCUMENT RESUME

ED 110 485 TM 004 758

AUTHOR Lee, Ann M.; Holley, Freda M.
TITLE An Ideal Evaluation Design in a Public School Setting: Or Where are You Campbell and Stanley Now That We Need You?
PUB DATE [Apr 75]
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Compensatory Education Programs; Elementary Secondary Education; Evaluation Methods; *Federal Programs; Hypothesis Testing; *Program Design; *Program Development; *Program Evaluation; Reading Improvement; Research Problems; School Districts; Teacher Aides

ABSTRACT

The first author set out to design and secure funding for an hypothesis-based program in a public school setting. The natural history of what happened to that study as it proceeded from design, to funding, to actual implementation, to final reporting serves as the case history of two idealistic evaluators' wildest nightmares. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED110485

AN IDEAL EVALUATION DESIGN IN A PUBLIC SCHOOL SETTING:

OR

WHERE ARE YOU CAMPBELL AND STANLEY NOW THAT WE NEED YOU?

Ann M. Lee, Ph.D. and Freda M. Holley, Ph.D.

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM 004 758

A Paper Presented at the Annual Conference of the American Educational Research Association

April, 1975, Washington, D.C.

The opinions expressed in this paper do not reflect the position or opinion of any organization or person other than of the authors themselves and no official endorsement should be inferred.

This paper is a detailed history of the implementation and evaluation of a federally funded compensatory program in a local school district. As is the case with most federal projects, the program design was revised on almost a day to day basis. The consequences of these revisions and associated events on the corresponding evaluation of the project are described in this paper. These events are discussed and several recommendations concerning future public school evaluation efforts are made.

The Emergency School Assistance Act (ESAA) was approved by Congress in 1972 to provide assistance to school districts involved in the desegregation process. Some of these ESAA funds were earmarked by Congress to finance pilot projects that would implement promising educational innovations. These funds were to provide needed compensatory educational aid and to finance the evaluation of these innovations in the hope that successful ideas would be replicated on larger scales.

When the district was notified of the availability of ESAA monies, a district evaluation unit was being established. Only coincidentally the person eventually responsible for heading up this evaluation unit also had a major role in designing the ESAA pilot project proposal. A local question (which also seemed to apply nationally) concerned the use of classroom aides in compensatory education programs and the concurrent effects on student behavior. Therefore, on the basis of local needs and federal guidelines, a pilot program was designed to test the following hypothesis:

> Students in classes with trained reading instructional aides will learn to read better than students in classes with untrained general aides and also better than students in classes with no aides at all.

The evaluator who was designing the program had recently read an article by Mssrs. Campbell and Erlenbacher[1] in which an elegantly stated case was made for "random assignment of children to treatments where this is possible" in compensatory educational programs. Persuaded by their rational arguments, the evaluator/designer decided to apply their suggestions to the program she was designing. Elementary schools designated as ESEA Title I constituted the target area for the project. Teacher volunteers for the project were to be solicited from the target area schools such that volunteer teachers could be randomly assigned to each of three groups: one group of teachers would receive the services of a trained reading instructional classroom aide; another group would receive the services of an untrained general aide; and the third group of teachers would receive no aide services at all. The resulting design was a pre-posttest randomized control group design described by Campbell and Stanley.[2] This design had the important advantage of eliminating school effects:

R	O_1	X_1	O_2
R	O_3	X_2	O_4
R	O_5		O_6

where X_1 = instructional reading aides
X_2 = untrained general aides

FIGURE 1

---

[1]Campbell, D. T., & Erlebacher, A. How Regression Artifacts in Quasi-Experimental Evaluations Can Mistakenly Make Compensatory Education Look Harmfull. In J. Hellmuth (Ed.), Disadvantaged Child - Compensatory Education: A National Debate (Vol. 3). New York: Brunner/Mazel Publishers, 1970.

[2]Campbell, D. T., & Stanley, J. C. Experimental and Quasi-experimental Designs for Research. Chicago: Rand McNally & Company, 1970. (Reprinted from Handbook of Research on Teaching, 1963.)

This design never made it off the drawing board. One objection to it was that the differentiation of aide services among teachers in the same building would not be tolerated by those teachers who were not also receiving the services of trained instructional reading aides. Another objection was that confining the program to only elementary schools would not promote instructional continuity from elementary to secondary school levels.

The design had to be altered in response to both of these objections. It was decided to abandon the randomized assignment of classes to treatment groups, and to instead assign complete schools to treatment groups. It was also decided to include a junior high school in the treatment group. Because of the limited number of aides provided by ESAA funds, the result of these design alterations was to reduce the number of schools in the treatment group from nineteen elementary schools to two elementary schools and one junior high school. This design a la Campbell and Stanley is represented below:

Elementary Level	O_1	X_1	O_2
	O_3	X_2	O_4
	O_5		O_6
Secondary Level	O_7	X_1	O_8
	O_9	X_2	O_{10}
	O_{11}		O_{12}

where X_1 = instructional reading aides
X_2^1 = untrained general aides

FIGURE 2

The two lowest-achieving elementary schools and the lowest-achieving junior high school in the district which needed compensatory services the most were selected as the experimental schools to receive the services of trained classroom reading instructional aides. Two Title I elementary schools which already had general classroom aides through the auspices of another federal compensatory program were selected as the elementary general aide comparison schools. The second lowest-achieving junior high school in the district was selected to receive the services of untrained general aides and to serve as the general aide secondary comparison school. The schools selected as the no-aide comparison group were two elementary schools which, unfortunately for the design, ranked in the top fourth, academically, of Title I schools in the district. The third lowest-achieving junior high school in the district was selected as the no aide secondary comparison school. (Standarized achievement test scores of students at this third junior high school were, however, approximately one full year higher than either the experimental or the general aide comparison junior high schools.)

The limitations of the design at this point seemed insurmountable: at the elementary level there were only two units of analysis and at the junior high level only one unit. In addition, the no aide comparison group was initially superior to both the experimental group and the general aide group. It appeared obvious to the evaluators even before the project had been implemented that there was little hope of answering the research question. Yet the evaluators hoped that some information could be gleaned during the next two years which might offer some clues to the most effective use of classroom aides. Little did they know what was in store.

Several unanticipated events occurred during the months just prior to the initial project implementation which affected the program design drastically. At the same time the ESAA Pilot funds for this project were awarded to the district, ESAA Bilingual/Bicultural funds were also awarded and were placed in the same three experimental schools as the pilot project. About one month prior to the opening of school, a court desegregation order required that sixth graders be moved from elementary buildings, and be bused to newly created sixth grade schools all over town. Two of these sixth grade schools were housed with the junior high experimental and general aide comparison schools. The effect this had on the project and comparison schools was to remove sixth graders from the elementary buildings and to incorporate them into the junior high school buildings, thereby altering the organizational and social structures at both levels. Then, approximately two weeks before school started, the two elementary project school principals were reassigned and two new principals, both young men in their first administrative assignment, were appointed.

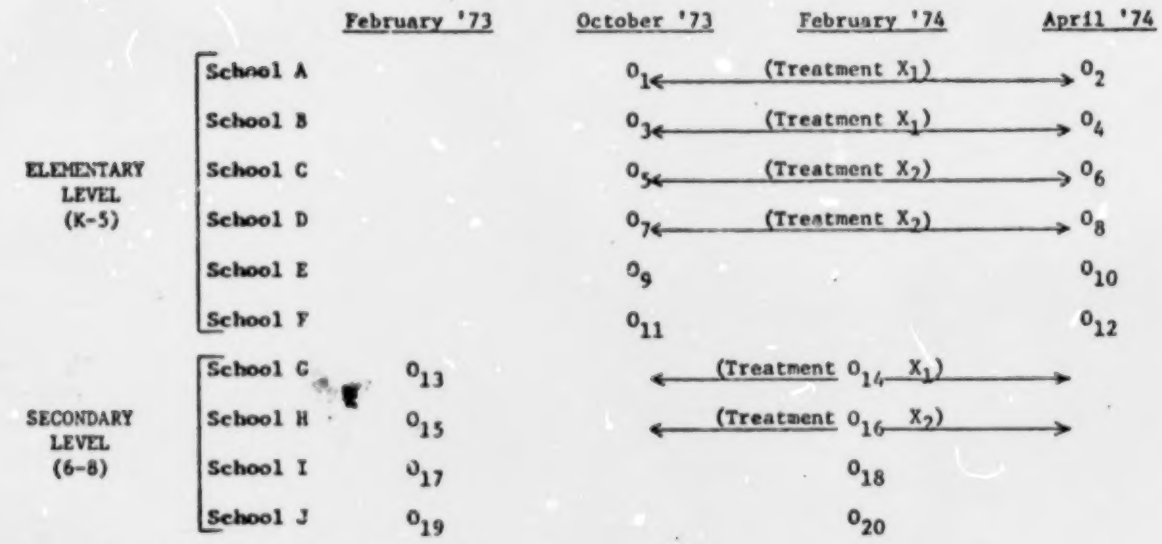
About two weeks after school started it became apparent that something unusual was going on in the elementary general aide comparison schools. Upon closer inspection, a special reading program sponsored by a local university was discovered by the evaluation staff to be operating in those comparison schools. This university project utilized approximately 80 part-time undergraduate tutors. We began to think that our experimental project schools were going to serve as a control group for the general aide comparison group. The design was getting more complicated:

ELEMENTARY LEVEL (K-5)	School A	O_1	X_1	X_3	X_4			X_7		O_2
	School B	O_3	X_1	X_3	X_4			X_7		O_4
	School C	O_5	X_2		X_4			X_7	X_8	O_6
	School D	O_7	X_2		X_4				X_8	O_8
	School E	O_9			X_4					O_{10}
	School F	O_{11}			X_4					O_{12}
SECONDARY LEVEL (6-8)	School G	O_{13}	X_1	X_3		X_5				O_{14}
	School H	O_{15}	X_2	X_3						O_{16}
	School I	O_{17}								O_{18}
	School J	O_{19}					X_6			O_{20}

where X_1 = instructional reading aides
X_2 = untrained general aides
X_3 = new bilingual program
X_4 = sixth graders removed from elementary buildings
X_5 = sixth graders introduced into junior high buildings
X_6 = new school for sixth graders only
X_7 = new first year principals
X_8 = special university reading project

FIGURE 3

As if matters were not already bad enough, problems were discovered with the project testing schedule. Because of understandable resistance from schools to over-testing of students, the project test measures were administered at pre-treatment and post-treatment times; some "pre" measures were given a half year before the start of the project, and the corresponding "post" measures were given halfway through the project year. This situation is represented in Figure 4 on the next page:



where X_1 = instructional reading aides
X_2 = untrained general aides

FIGURE 4

During the second project year other special programs were introduced into the project and comparison groups, leaving the "design" looking something like Figure 5 below:

	School A	O_1	X_1	X_3	X_4		X_7			X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	O_2
	School B	O_3	X_1	X_3	X_4		X_7			X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	O_4
ELEMENTARY LEVEL (K-5)	School C	O_5	X_2		X_4		X_7	X_8	X_9							O_6
	School D	O_7	X_2		X_4			X_8	X_9		X_{11}					O_8
	School E	O_9			X_4				X_9							O_{10}
	School F	O_{11}			X_4				X_9							O_{12}
	School G	O_{13}	X_1	X_3		X_5					X_{11}				X_{15}	O_{14}
SECONDARY LEVEL (6-8)	School H	O_{15}	X_2	X_3												O_{16}
	School I	O_{17}							X_9							O_{18}
	School J	O_{19}				X_6			X_9							O_{20}

where X_1 = instructional reading aides
X_2 = untrained general aides
X_3 = new bilingual program
X_4 = sixth graders removed from elementary buildings
X_5 = sixth graders introduced into junior high buildings
X_6 = new school for sixth graders only
X_7 = new first year principals
X_8 = special university reading program
X_9 = another bilingual program
X_{10} = Teacher Corps training program
X_{11} = behavior modification training program
X_{12} = social studies curriculum pilot project
X_{13} = new reading curriculum
X_{14} = social workers
X_{15} = university tutors

FIGURE 5

DISCUSSION

It is not the point of this paper to belittle our own efforts nor the efforts of other evaluators. Nor do we disagree in theory with any criticisms of quasi-experimental and ex post facto evaluation designs. However, warnings against regression artifacts and matched samples seem somewhat irrelevant when applied to the real-life problems described previously in this paper. We feel that more fundamental warnings are needed for today's evaluators and educational decision-makers.

When programs are selected for evaluation (almost always after the program design is completed) they are assigned to either an internal or an external agency for evaluation. Usually, however, the program has been designed so poorly (for evaluation purposes) that very little can be discovered concerning its worth. These design inadequacies most often result from very real political pressures brought to bear upon decision makers: they are urged to blanket a whole population with the latest curriculum (reserving no valid subjects for control purposes) or to introduce all at once literally dozens of resources into a few schools (thereby concealing the relationship between individual treatments and outcomes). In both cases, finding answers to crucial evaluation questions is almost guaranteed to be impossible. In addition to being a frustrating situation for evaluators and their bosses, an evaluation of such programs under these conditions does not yield a maximum return on taxpayers' money.

Evaluation is still an infant in the education family. Persons not directly involved in it have developed little appreciation for those events which can invalidate evaluation conclusions. We would like to emphasize here a few of the basic rules of program design which must be adhered to if needed answers are to be provided through evaluation:

BASIC RULES FOR PROGRAM DESIGNERS

1. When the merits of an educational program are to be assessed, the treatment group must be compared to some control group in which the treatment is not present. Otherwise, any gains or losses observed among the treatment group cannot be unequivocally attributed to the treatment which is being evaluated.

2. The treatment and comparison groups must be composed of the same kind of people. Random assignment of subjects to each group is the most reliable way to attain identity of groups, although matching can be used if a large enough subject pool is available. If matching is used, it must be done on a large number of variables and not on just a few.

3. There must be a large enough number of units in each group to allow for the plausibility of significant differences occurring between them, and to allow for any degree of generalizability of the results. We might point out here that if 500 students in two schools are assigned to a treatment group and 500 students in two other schools are assigned to a control group, the number of statistical units in each group is two, not 500. Usually the differences among small groups of schools due to non-treatment sources like socioeconomic status, staff competencies, etc., are so many that any differences between the schools due to the treatment are obscured. If a large number of schools is not available for assignment to the treatment and control groups, then the treatment should be randomly assigned on either a classroom or an individual student basis, as appropriate.

4. All subjects in both the treatment and control groups must be pre-tested at the same time and post-tested at the same time. (A test administration which covers a one-month period or longer does not qualify as being "at the same time.")

5. Treatments should not be compounded in either the experimental or the control groups. If one curriculum is being compared with another curriculum, other large-scale programs should not also be distributed among the treatment and control groups.

These rules of program design are certainly very basic ones which are so familiar as to probably insult most of our readers. However, we believe that there are few educational program designs being implemented and evaluated today which do not violate several of these obvious rules. We think the main source of problems with program designs is that funding agencies, local education agencies, and educators in general do not understand the requirements for determining whether or not a program is successful. Nor do they understand the implications of not meeting these requirements. In order to reduce this lack of understanding, we would make the following recommendations:

RECOMMENDATIONS TOWARD IMPROVING PROGRAM DESIGNS

1. Evaluators in local education agencies must initiate or step up their inservice efforts with decision-makers concerning design requirements of programs which are to be evaluated. This training should involve school board members, district-wide administrators and principals at the very least.
2. All preservice teacher training programs should include a required introductory course in educational research and evaluation design and methodology. This would yield classroom dividends over and above benefits to those educational programs in which teachers would eventually participate.

3. All administrator certification programs should require both introductory and advanced courses (at least six hours total) in educational research and evaluation design and methodology.

4. The educational research community should promote and sponsor conferences for educators who are not directly involved in research or evaluation, but who are responsible for program planning and design. The purpose of these conferences would be to communicate research and evaluation requirements in program design. For example, the American Educational Research Association could develop and sponsor these conferences in conjunction with the American School Board Association and the American Association of Public School Administrators.

5. Public school evaluation units should have approval authority on the design of those programs which are to be evaluated. If evaluators do not have this authority, their predicament can become a question of ethics as described in the following situation:

   A public school evaluation unit is directed to find out if Curriculum X has a beneficial effect on student achievement. But the evaluation staff realizes that the program design set up by the instructional department and approved by the school board will prohibit this question from being clearly answered. There are no control groups established for comparison with the treatment group, or if control groups have been established, the two groups are nowhere near identical; other independent variables (Curriculums A, B, C, D, E, F, G and H) are so liberally distributed throughout both the Curriculum X group and the "control" group as to make any achievement gains in either group totally uninterpretable with respect to Curriculum X. Should the evaluation unit go ahead and "evaluate" the program as it is designed, or should they refuse to do so on ethical grounds?

The above example clearly illustrates that what we must do is educate our colleagues (who do have educational planning and budgeting responsibilities) concerning the committments they must make if they really want useful evaluation information. We need to be hardnosed and persistent in these efforts, and perhaps even decline to evaluate an impossibly designed program.

As evaluators we are also accountable, and the measure of our effectiveness is the improvement observed in student learning. This improvement will not occur if the information we provide decision-makers is invalid or if it is not used as input into the decision-making process. Our frank conclusion is that until our program designs improve, and until our colleagues are taught how to use the evaluation data we provide, rational decision-making and accountability will not occur.